CSC 440 HW 1 Report LinxiaoBai

Book Question

1.1

Data mining is in fact the method and process of mining knowledge. The term involves greater attention because internet and cloud database accumulated and stored much greater amount of data. More advanced computation engine allows heavier work load and process capacity. The direct result is more and more knowledge are allowed to be mined at faster speed.

Data mining is not an advertising, but it is closely related to computational advertising and online recommendation system. For example, data manage platform relies mostly on data mining to label and classify its customer.

The technique of data mining not only involves database management, statistics model, machine learning, pattern recognition. Neither is the simple combination of the knowledge above. It includes the understanding of the data, and the pattern of the world.

The general procedure of data mining usually starts with the step of accessing data. Then followed by data cleaning, transformation. Then if the mining objective is clear, say discover certain correlation between features, it could start with correlation studies. Also if labeled observations are provided, supervised data mining can be employed. Alternatively, data mining can rely on rule based, and unsupervised learning techniques.

1.2

The major difference is data warehouse is designed for long term and resilient storage.
It means that data is meant to store for long period and long term analysis.
The storage format is usually with row-based-storage high compression rate, low R/W speed in order to save drive and relieve IO burden.

However, database is designed to be store recent used for fast query and analysis. The storage is usually column-based, with low compression rate, and high R/W speed for fast query.

Both data warehouse and database are storage for data, and they are inner transferable. Without consideration of their effectiveness toward certain task, they are replaceable to each other.

1.4
Alibaba and its recommendation system used the habit analysis result and recommendation algorithm of their DMP department. The result and algorithm is based on analysis and data mining of their customer's purchase and browse result.
The directed recommendation strategy increases the company's 11.11(Chinese Black Friday) income by at least 30%.
1.5
Discrimination VS Classification

The former refers a general comparison of specific features between current class and target class. While the later refers to the process of determine a class among a selection of candidates based on known information and algorithm.
The later meaning carries a more specific task.

Characterization VS Clustering
The former is a process that estimate and evaluate the determinant feature of certain class.
The later is a method of unsupervised learning.

Classification VS Prediction
The former is a process of classify an unknown observation to a limited set of selection of classes.
The later is very similar to the first one except the out put is a real number between prediction domain. Not necessarily a value of a set.

1.7
Strategy 1:
simulate/ guess the distribution of certain features. Then based on statistics test to if certain observation falls in the range of the outliers based on certain significant level.

Strategy 2:
Cluster analysis/ unsupervised learning. Find points out of clusters.

Personally, I think the later one is more reasonable. Because the first one strategy relies on guessing the distribution of the variables. This certain step might be very inaccurate, and causes misunderstanding of the data.

1.9

The major challenge is the computation burden and running time.

Computation burden is represented as as data grow bigger, it consumes bigger memory to process. The relationship of memory, CPU resource required and the size of the data could be, determined by the data mining algorithm, not only linear but also exponential. This requires bigger memory, more and faster CPUs, bigger distributed file system, and faster distributed computation engine.

In terms of Run-time it refers to the growing need of online, real-time computation targeted for specific task. This requires faster algorithm, and better optimization. Usually, the training time of a certain algorithm should no more than O(klog(n)).

None textbook part

Data Chosen:

Auto.csv from ISLR library of CRAN

Format:

A data frame with 392 observations on the following 9 variables.

mpg miles per gallon

cylinders Number of cylinders between 4 and 8 displacement Engine displacement (cu. inches) horsepower

Engine horsepower

weight Vehicle weight (lbs.)

acceleration Time to accelerate from 0 to 60 mph (sec.) year Model year (modulo 100)

origin Origin of car (1. American, 2. European, 3. Japanese) name Vehicle name

The original data contained 408 observations but 16 observations with missing values were removed *((2013) Package 'ISLR.' Package 'ISLR')*.

Running basic statistics :

a)

	mpg	cylinders	displacement	weight	acceleration	year	origin
count	397	397	397	397	397	397	397
mean	23.51586902	5.458438287	193.5327456	2970.261965	15.55566751	75.99496222	1.574307305
std	7.825803929	1.701576981	104.3795833	847.9041195	2.749995293	3.690004901	0.802549496
min	9	3	68	1613	8	70	1
25%	17.5	4	104	2223	13.8	73	1
50%	23	4	146	2800	15.5	76	1
75%	29	8	262	3609	17.1	79	2
max	46.6	8	455	5140	24.8	82	3

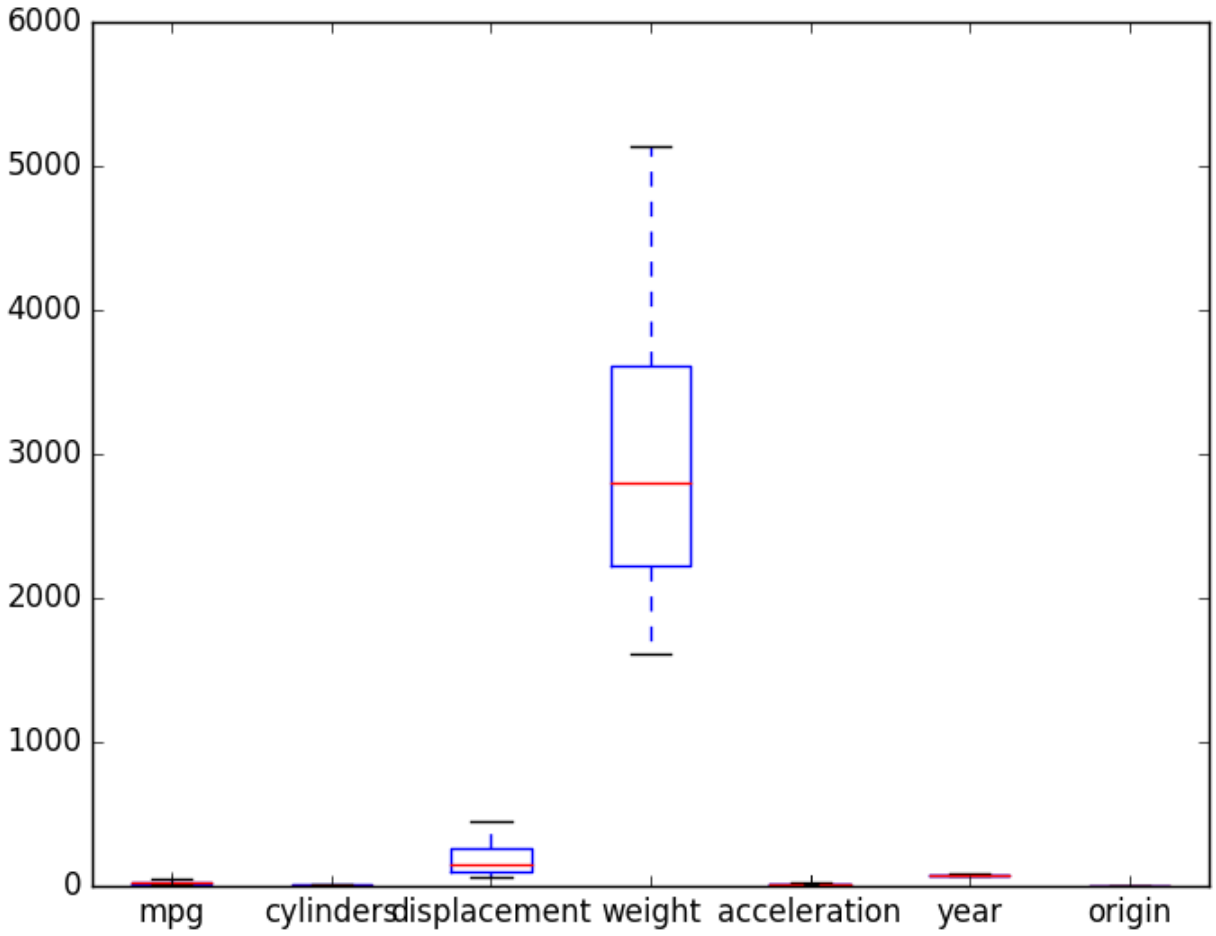
Analysis:
According to the basic statistics of the data, we see that the mean mpg of cars is 23.5 very close to its median. This suggest that 23 is a good estimator of the overall mpg status.

Also, most of the cars run on 4 cylinders, with 3t of weight. Notice that the standard deviation is considerably low for cylinders. This suggest that the distribution of cylinders' number are pretty concentrated.

Acceleration has mean 15.5 and correspond to its median. This is a indication that the variable may not be screwed.

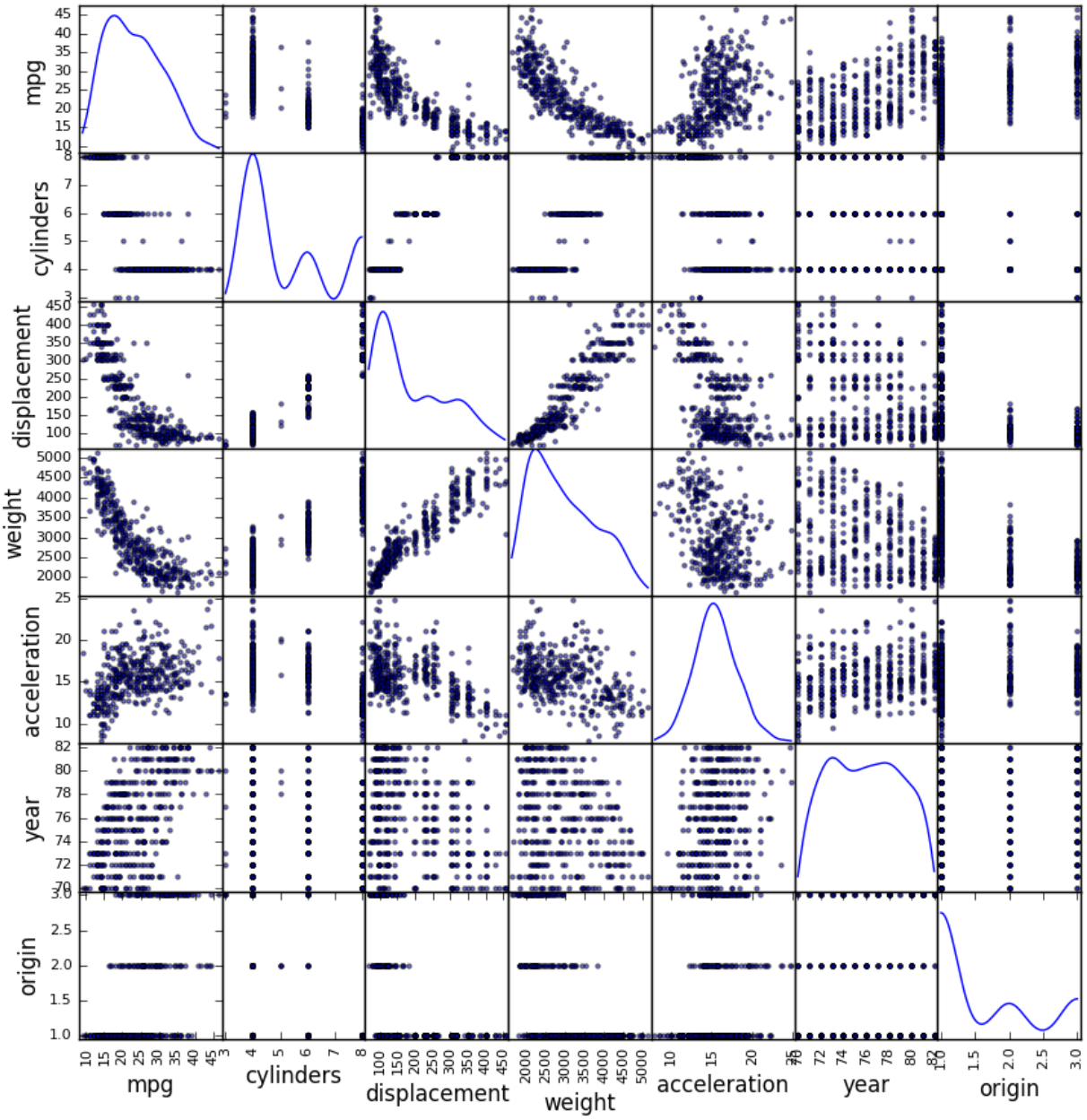
Origin is not very considerable value, because it is not numerical. The number is just coding of origin country.

b)



The boxplot reflects the general pattern of distribution and skewness of the data.

c)
Then we can plot scatter plot the analyze the dependency between variables.



By looking at the plot, we are almost certain that some variables are closely related. For example, Acceleration, cylinders, displacement, weight, and mpg are pair-wised related. Year and mpg are also related.

To verify the relation, we need to compute the correlation matrix.

mpg	1	-0.776259948	-0.804443045	-0.831738914	0.422297414	0.581469459	0.5636979(
cylinders	-0.776259948	1	0.950919865	0.897016868	-0.504060631	-0.346717221	-0.56497
displacement	-0.804443045	0.950919865	1	0.933104417	-0.544161823	-0.369804092	-0.6106643
weight	-0.831738914	0.897016868	0.933104417	1	-0.419502328	-0.307900414	-0.5812651(
acceleration	0.422297414	-0.504060631	-0.544161823	-0.419502328	1	0.282900891	0.2100836
year	0.581469459	-0.346717221	-0.369804092	-0.307900414	0.282900891	1	0.1843140
origin	0.563697905	-0.5649716	-0.610664323	-0.581265169	0.210083616	0.184314075	

The result shows that there indeed exists correlation between variables. For example, cylinders and displacement has correlation of .95 which is a very strong linear correlation. Also same discovery happens between cylinders and weight, and so on.